IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Arthur Thouvenin
03/09/2021

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- During this IBM-Coursera course we used a lot of methodologies such as :
  - Data Collection with APIs, web scrapping, SQL connections
  - Data Wrangling with Pandas, Numpy, SQL
  - Data Visualization with Plotly, Seaborn, Dash, Folium
  - Predictive Analysis with Scikit-Learn
- Summary of all results
  - Thanks to all of this projects we were able to have a better understanding of the success rate of space launch.
  - But also to find the launching pads in the US.

Introduction

- The goal of this project was to use every side of data sciences to study the space launch companies and especially the Falcon 9 rocket launch.

- First determine the price of each launch, but also gathering information about SpaceX and creating dashboards. If SpaceX will reuse the first stage of Falcon9 rocket, if it will land successfully thanks to machine learning.

Section 1

Methodology

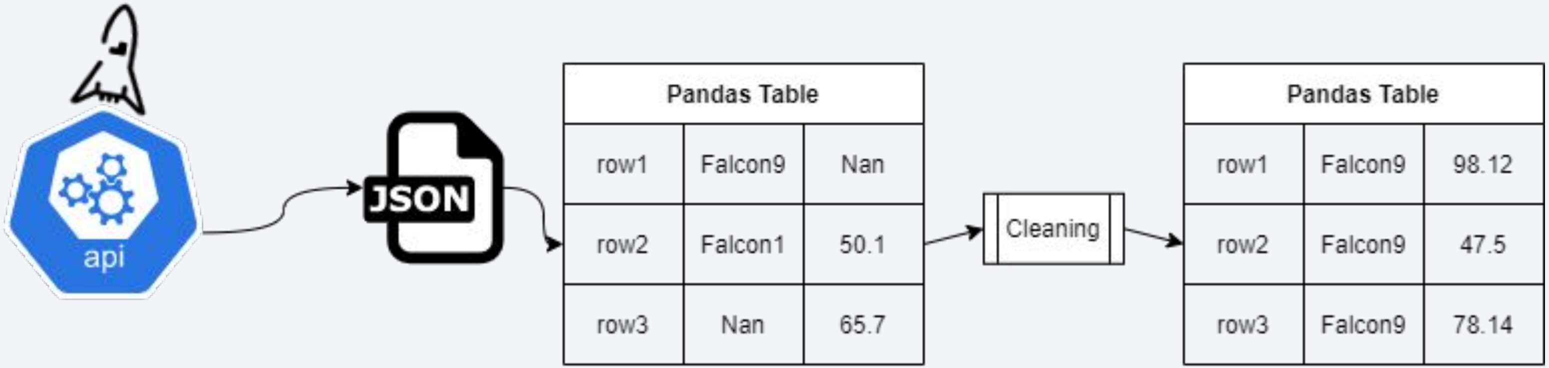
Methodology

Executive Summary

- Data collection methodology:
  - We used several methods from web scrapping, to SQL queries and APIs
- Perform data wrangling
  - The data were most of the time standaradize and processed with NUmpy and Pandas
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models were build using scikit-learn, we used a GridSearch with 10 fold of cross validation, in the end we use the accuracy to determine the best classifier

Data Collection

- We used the spaceX data API for data collection
- Thank to the spaceX data API we were able to collect launches data in the JSON format and then transform it into a Pandas DataFrame. Thanks to this DataFrame we were able to clean the dataset of missing values and rockets that we were not interested in such as Falcon 1



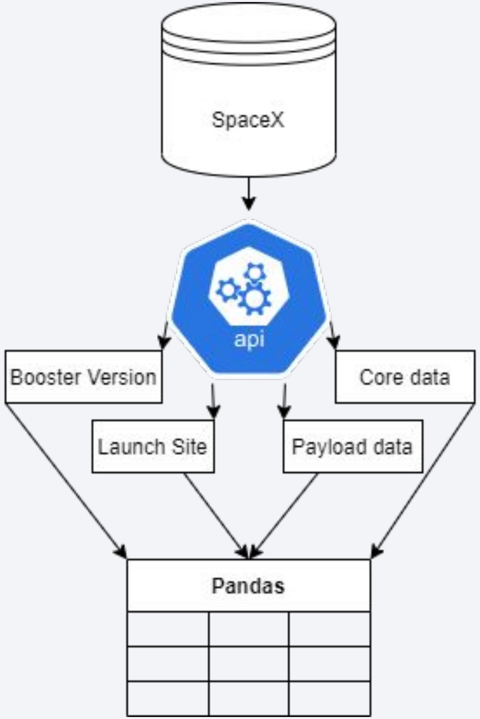
Pandas Table		
row1	Falcon9	Nan
row2	Falcon1	50.1
row3	Nan	65.7

Pandas Table		
row1	Falcon9	98.12
row2	Falcon9	47.5
row3	Falcon9	78.14

Data Collection – SpaceX API

- Thanks to multiple requests we were able to collect the Booster Version, the Launch Site, the Payload data and the Core data

- GitHub URL

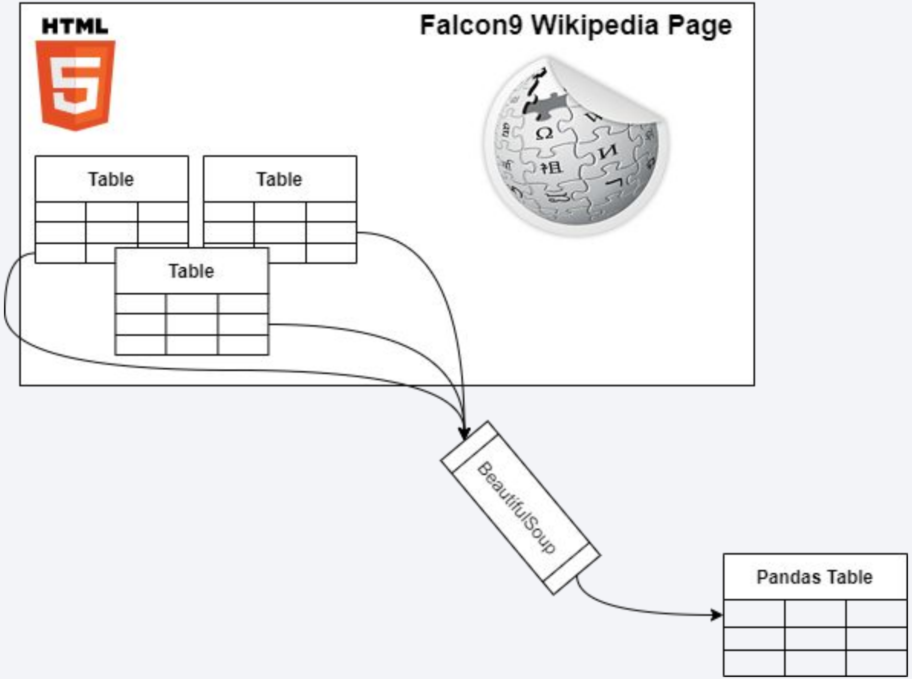


Data Collection - Scraping

- Thanks to *BeautifulSoup* it is easy to do web scrapping and especially collect tables from *HTML*.

- Here we used BeautifulSoup to extract the header of tables and then populate the table with the rows collected

- [GitHub URL]()

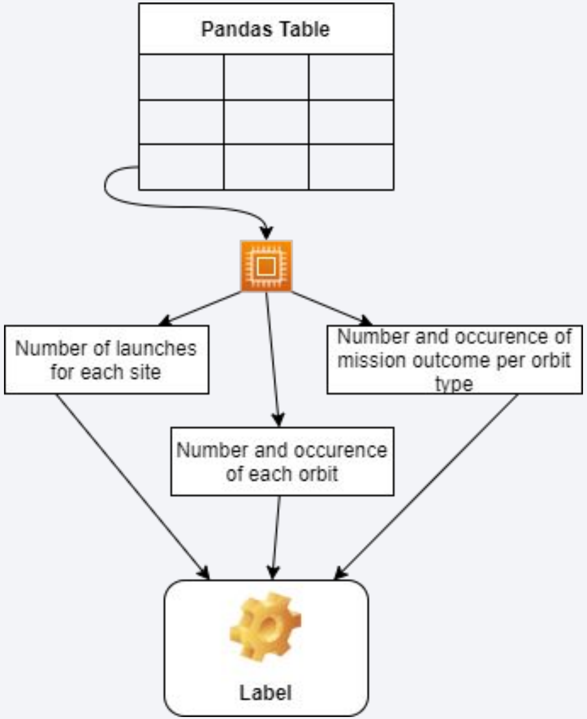


Data Wrangling

- Data were analyzed using Exploratory data Analysis, thanks to this we were able to determine the training labels.

- We computed the number of launches on each site, the number of occurrence of each orbit and the number and occurrence of mission outcome per orbit type

- [GitHub URL]()


Pandas Table
Number of launches for each site
Number and occurence of mission outcome per orbit type
Number and occurence of each orbit
Label

EDA with Data Visualization

- We used :
  - A scatter plot of *FlightNumber* vs. *PayloadMass* based on the outcome to see if as the flight number increases, the first stage is more likely to land successfully and if there is an impact of PayloadMass
  - Then a second scatter plot of *LaunchSite* vs. FlightNumber based on the outcome to understand if the *LaunchSite* does not become a problematic element following the number of rockets which are launched there (*FlightNumber*).
  - A third scatter plot to observe if there is any relationship between *LaunchSite* and their *PayloadMass*
  - A bar chart to visually check if there are any relationship between the *Success Rate* and the *Orbit Type*
  - Another scatter plot to visualize the relationship between *FlightNumber* and *Orbit type*
  - A last scatter plot to reveal the relationship between *Payload* and *Orbit type*
  - And a line plot to get the average launch success trend
- GitHub URL

EDA with SQL

- SQL queries were performed as followed :
    - The names of the unique launch sites in the space mission
    - 5 records where launch sites begin with the string 'CCA'
    - The total payload mass carried by boosters launched by NASA (CRS)
    - The average payload mass carried by booster version F9 v1.1
    - The date when the first successful landing outcome in ground pad was achieved
    - Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - Total number of successful and failure mission outcomes
    - Names of the booster_versions which have carried the maximum payload mass
    - Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
    - The count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- GitHub URL

Build an Interactive Map with Folium

- In this part we marked the NASA Johnson Space Center with a blue circle and a Marker. Then we marked with Circle and Marker each launch which results in marking launching sites. Thanks to Marker we have marked successful and failed launches with green and red colors. In the end we add a Marker and a Line to the closest railway.
- All of those Markers Lines and Icons helped us understand the geography of launch sites and their proximity to cities, railways, coast etc..
- GitHub URL

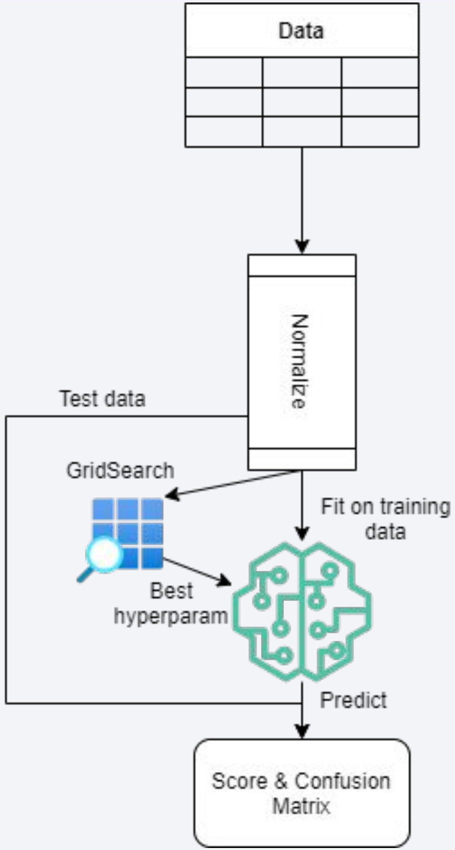
Build a Dashboard with Plotly Dash

- In the Dashboard we add a pie chart of success rate for each launch site and also a scatterplot of the payload mass by the outcome of the mission (class) all this based on Booster versions
- The first plot helps to understand the success rate on each launching site, the second demonstrate which booster as the higher success rate and that it is related to the Payload mass.
- GitHub URL

Predictive Analysis (Classification)

- Models were built using Scikit-Learn, data were previously normalized and models hyperparameters were found using a GridSearch with a 10 fold cross validation, in the end the best performing model has been selected based on accuracy.

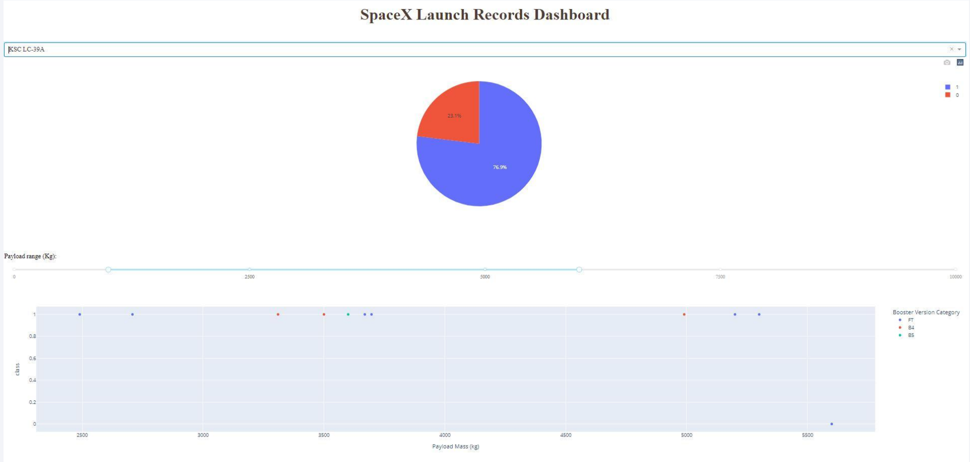
- GitHub URL



Results

- The EDA demonstrate the importance of the payload mass, but also the booster version on the success rate of the launch missions

- All machine learning models studied were equally performing



SpaceX Launch Records Dashboard

KSC LC-39A

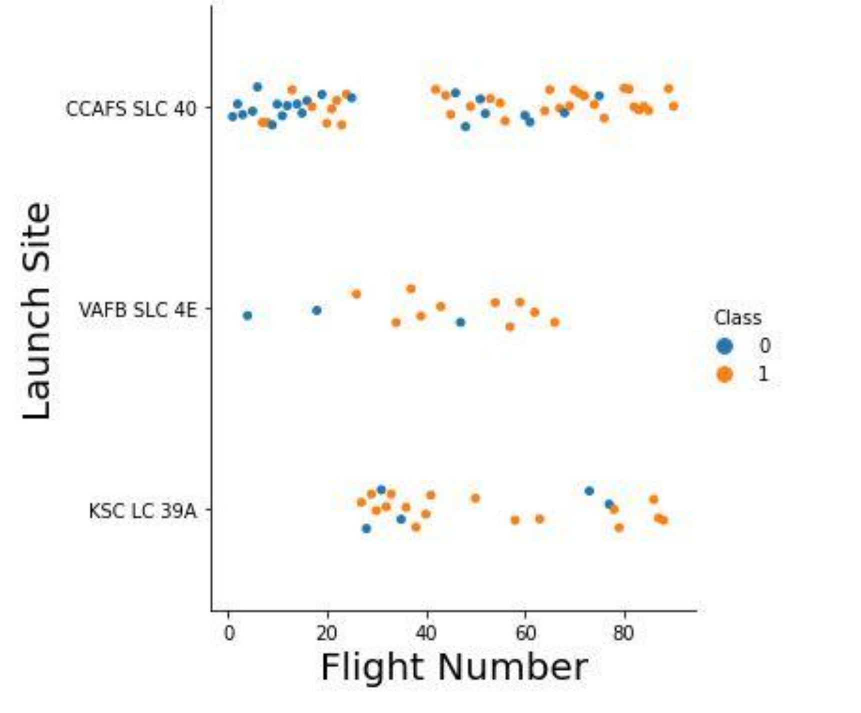
Payload range (Kg):

Section 2

Insights drawn from EDA

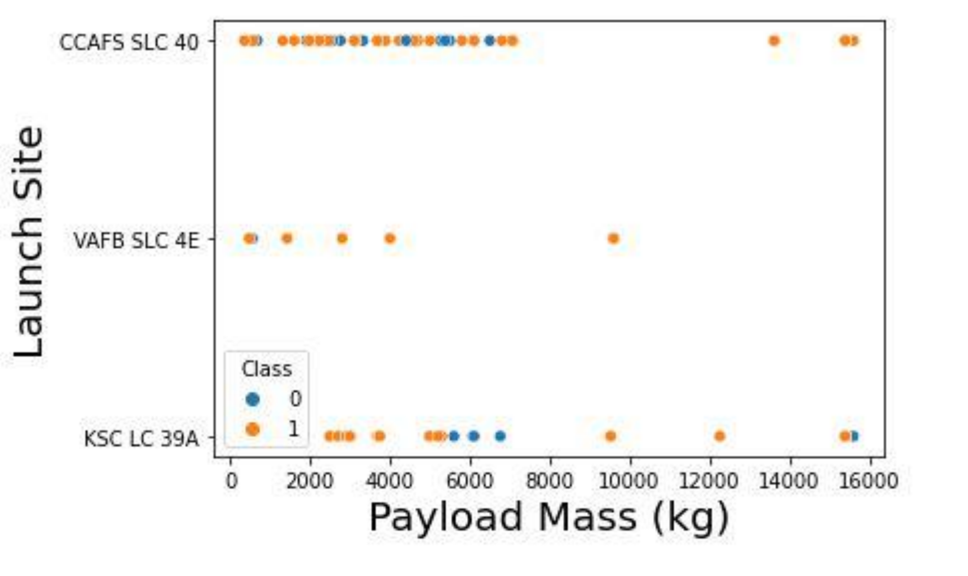
Flight Number vs. Launch Site

This chart seems to indicate that a “young” launching site will probably a lower success rate than one which had a lot of rocket launched from.



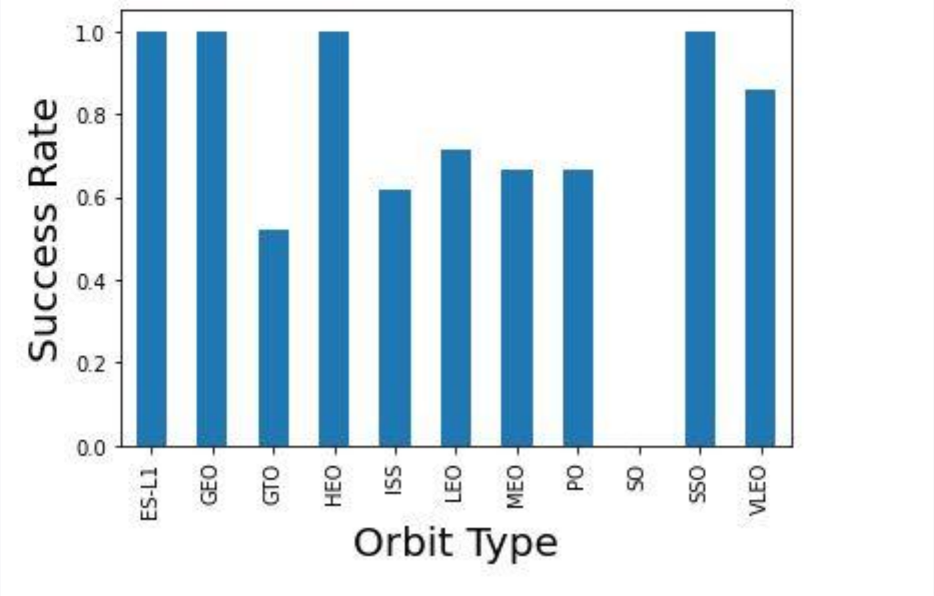
Payload vs. Launch Site

It seems like a lot of rocket launched had a payload between 500kg and 6000kg. Also the launching site VAFB SLC 4E seems to be a site where there are not that much rocket launched. An impact of the payload could be possible but it will need further analysis.



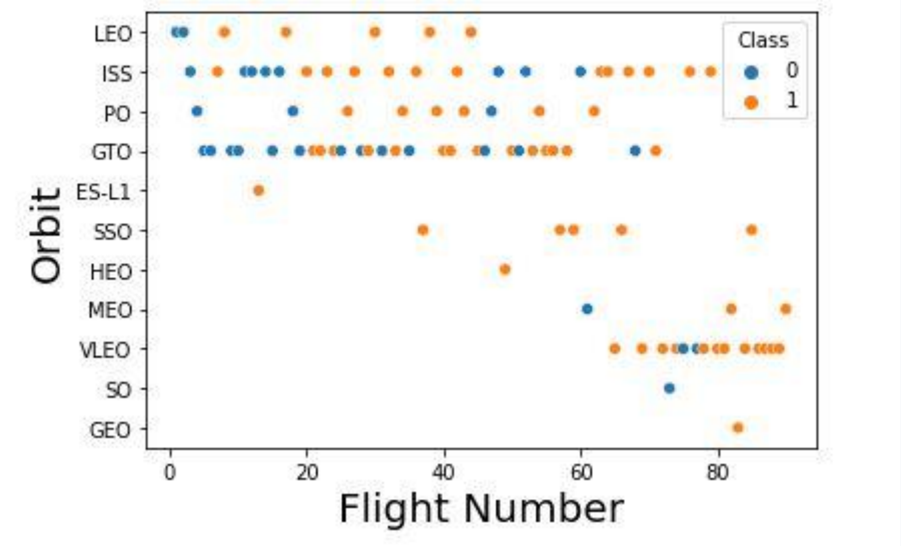
Success Rate vs. Orbit Type

There is a strong correlation between these two indeed as we can observe the SO or GTO Orbit Type are quite risky as the success rate is below 0.6. However, some Orbit Type provide a 1.0 success rate which is perfect but can hide suspicious data. Indeed if for this Orbit type only one rocket has been launched the reliability of this hypothesis is null.



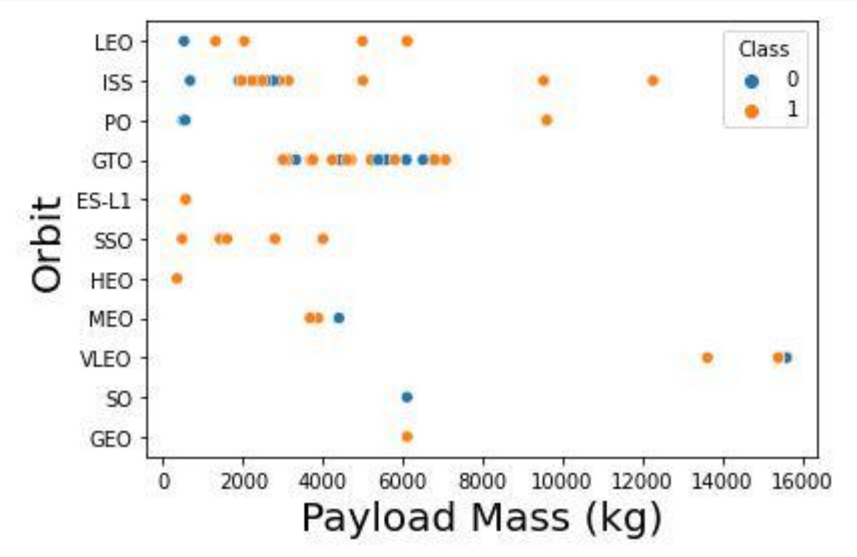
Flight Number vs. Orbit Type

This chart confirmed what has been said before, some Orbit type have only couples of Flights in their history and thus make data quite confusing. However for the GTO,VLEO and ISS it seems like there are enough data to be confident on those data.



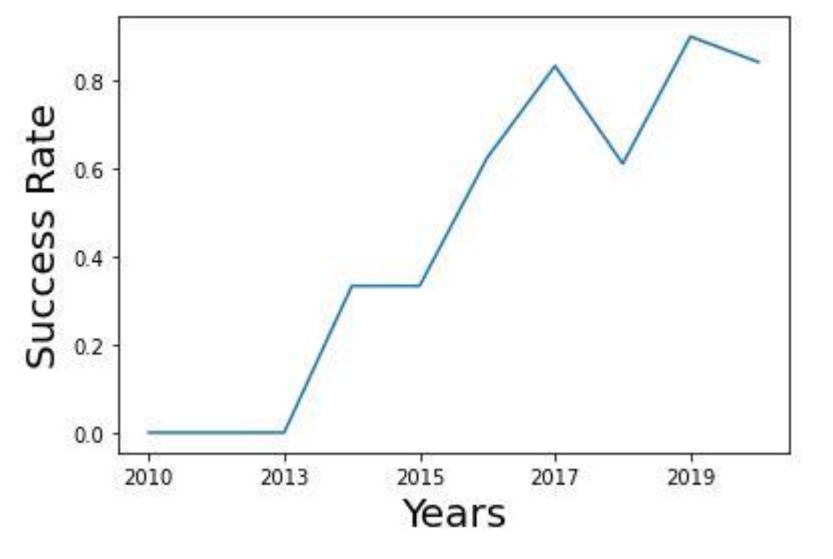
Payload vs. Orbit Type

Here we can observe that certain sites have a strong relation with the payload mass, for example the GTO and ISS.



Launch Success Yearly Trend

Here the chart demonstrates that as Humans learn more and more through the years thanks of Sciences,it results in a significant rocket launches success rate increasing.



All Launch Site Names

- Launch sites are :
    - CCAFS LC-40
    - CCAFS SLC-40
    - KSC LC-39A
    - VAFB SLC-4E
- Those has been collected thanks SQL query

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` :
    - CCAFS LC-40
    - CCAFS LC-40
    - CCAFS LC-40
    - CCAFS LC-40
    - CCAFS LC-40

- Those has been collected thanks SQL query

Total Payload Mass

- The total payload carried by boosters from NASA is 99.980 kg

- This number has been collected thanks SQL query

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is :
    - 2.534 kg

- This number has been collected thanks SQL query

First Successful Ground Landing Date

- The dates of the first successful landing outcome on ground pad was :
  - 2015-12-22

- It has been collected thanks SQL query

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 :

  - F9 FT B1032.1
  - F9 B4 B1040.1
  - F9 B4 B1043.1

- Those has been collected thanks SQL query

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes :
  - Those has been collected thanks SQL query

Mission Outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List of the names of the booster which have carried the maximum payload mass :
  - F9 B5 B1048.4
  - F9 B5 B1049.4
  - F9 B5 B1051.3
  - F9 B5 B1056.4
  - F9 B5 B1048.5
  - F9 B5 B1051.4
  - F9 B5 B1049.5
  - F9 B5 B1060.2
  - F9 B5 B1058.3
  - F9 B5 B1051.6
  - F9 B5 B1060.3
  - F9 B5 B1049.7
- Those has been collected thanks SQL query

2015 Launch Records

- List of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Landing Outcome	Booster version	Launch Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- Those has been collected thanks SQL query

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

- Present your query result with a short explanation here

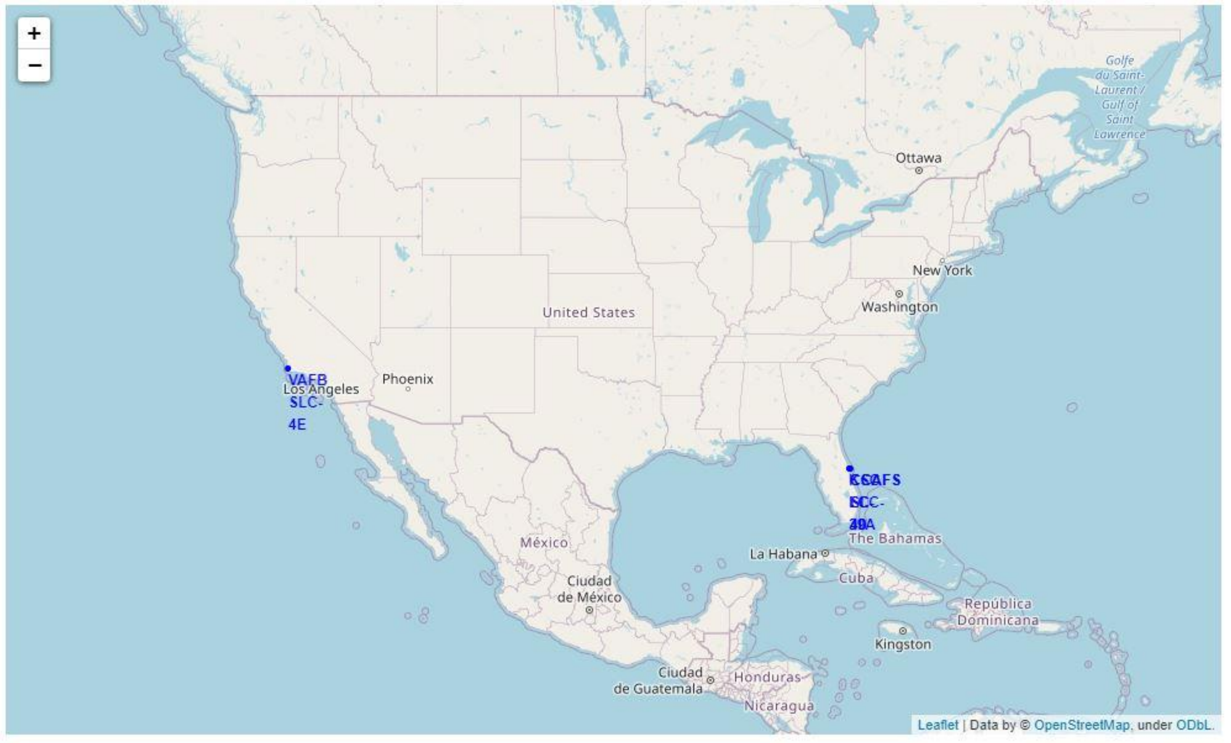
Landing Outcome	Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

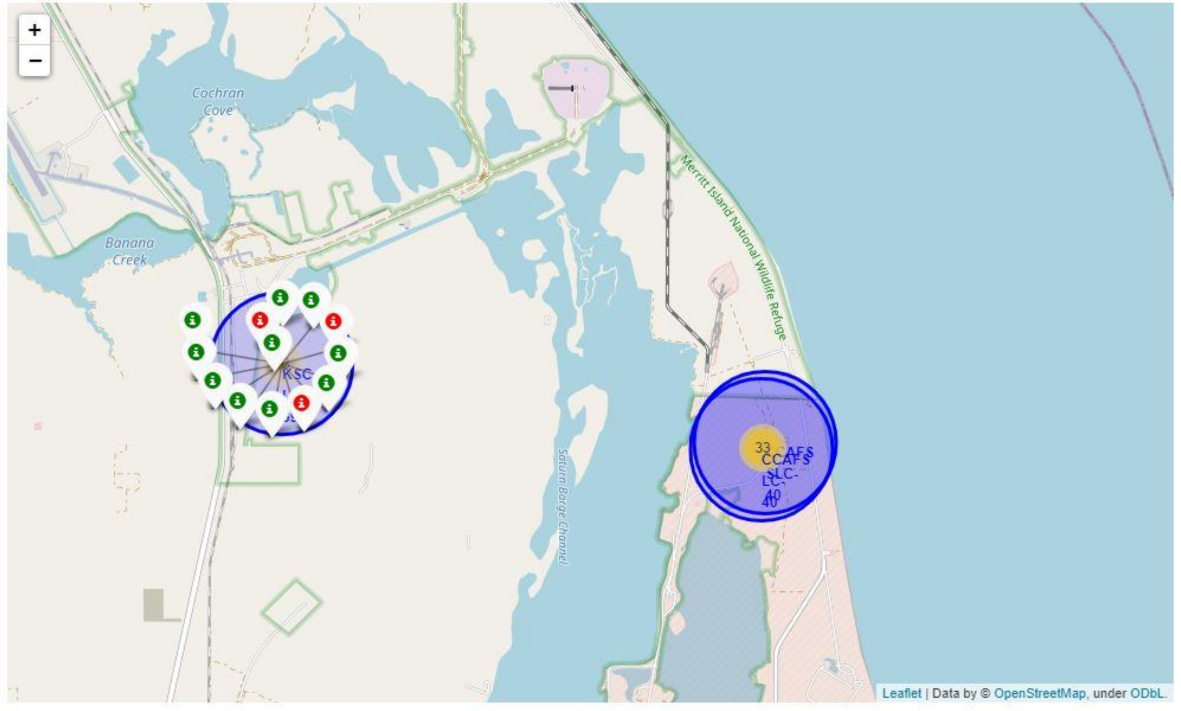
Launch sites in the US

Here we can observe launching sites in the US marked in blue, it is a little bit hard to see in Florida as the three sites are very close.



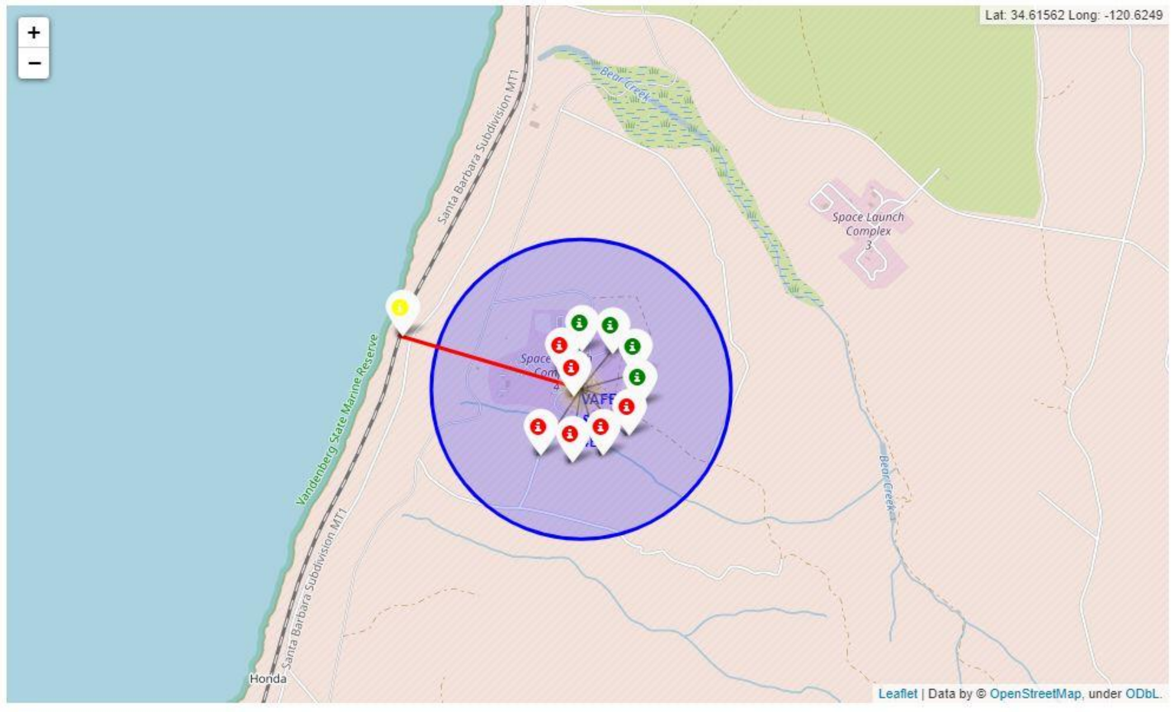
Success or Failure Marker

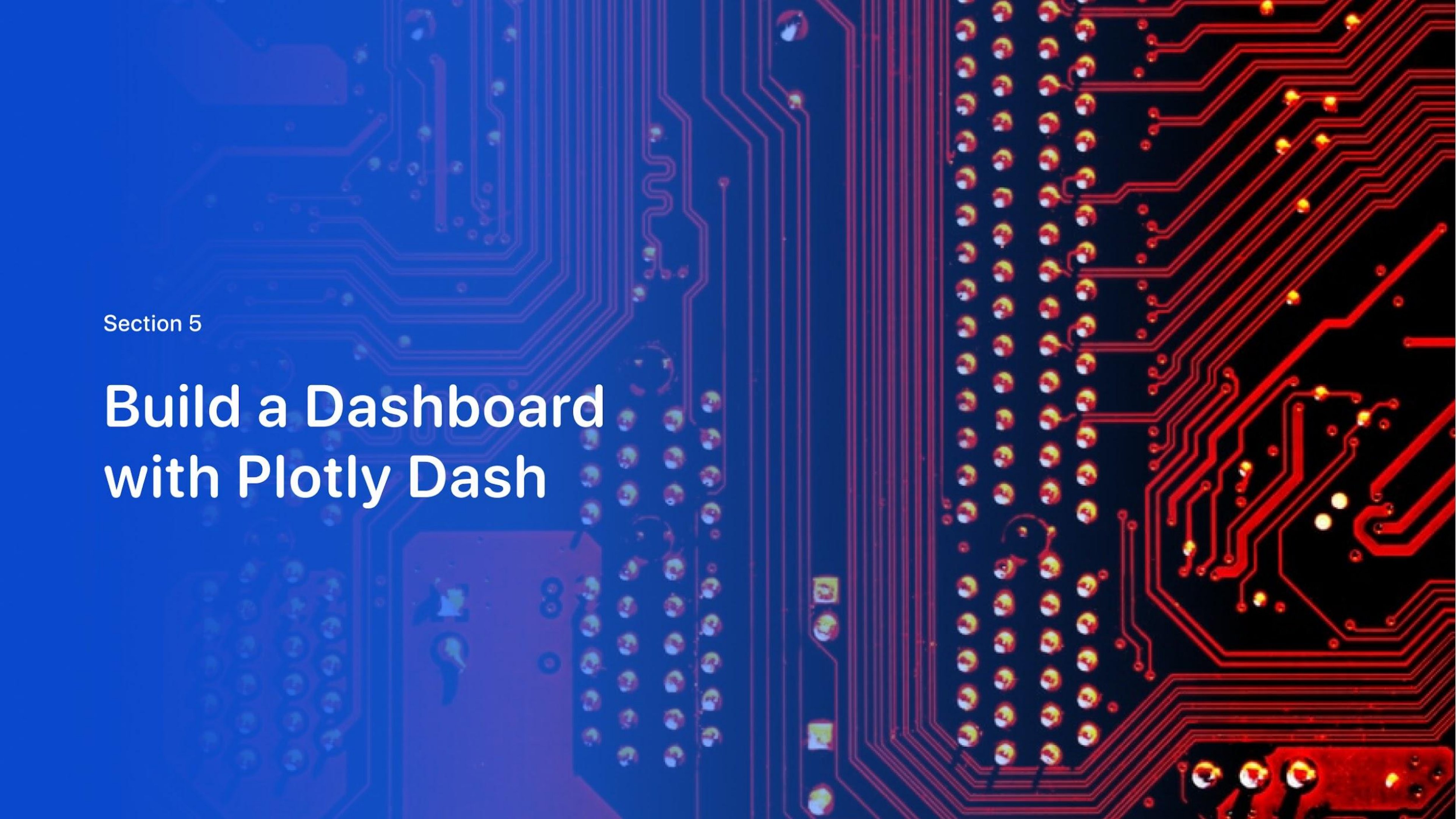
Here we have, the two overview of these icons, one with two sites overlapping one another, and one with all the icon success (green) and failure (red)


+
−
Cochran Cove
Banana Creek
Merritt Island National Wildlife Refuge
Saturn Barge Channel
KSC
33
CCAFS
AFS
SLC-
LC-
40
40
Leaflet | Data by © OpenStreetMap, under ODbL.

Launch site proximities

Here we can observe the proximity of the launching site and the trainline marked with a yellow icon and red line, but also its proximity with the coast

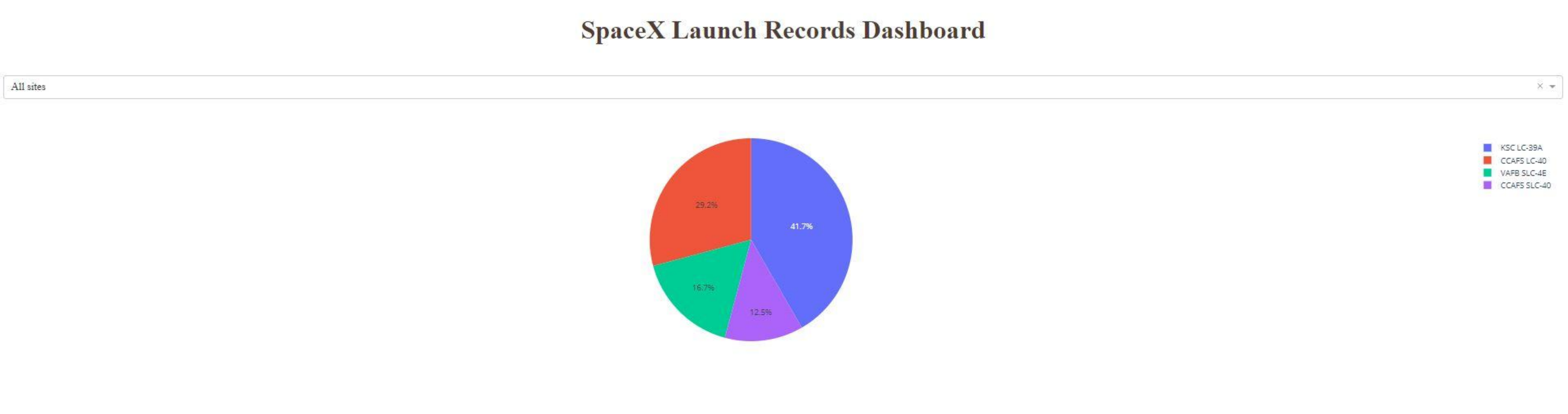

+
−
Lat: 34.61562 Long: -120.6249
Bear Creek
Santa Barbara Subdivision MT1
Space Launch Complex 3
Vandenberg State Marine Reserve
Bear Creek
Santa Barbara Subdivision MT1
Honda
Leaflet | Data by © OpenStreetMap, under ODbL.



Section 5

Build a Dashboard with Plotly Dash

Success rate of all sites

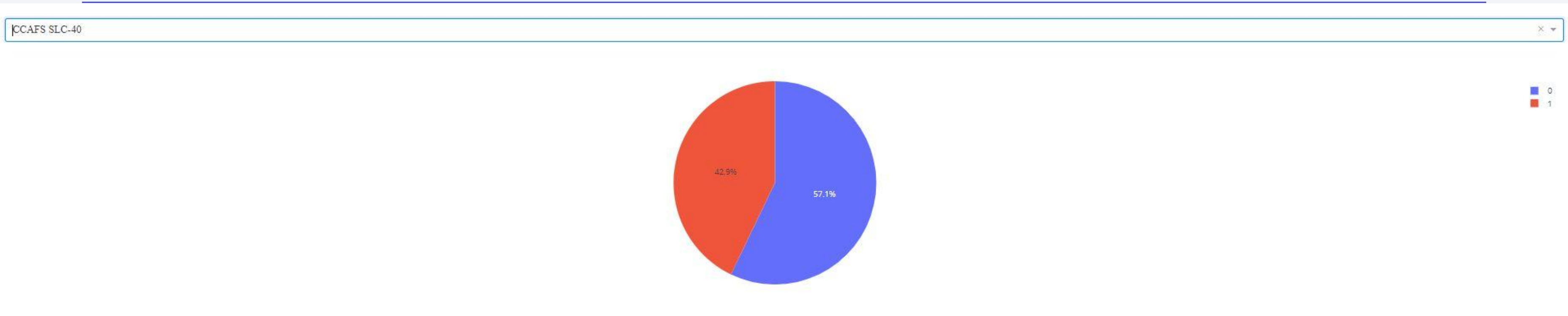


SpaceX Launch Records Dashboard

All sites

Here we can observe the different success rate for each launching site

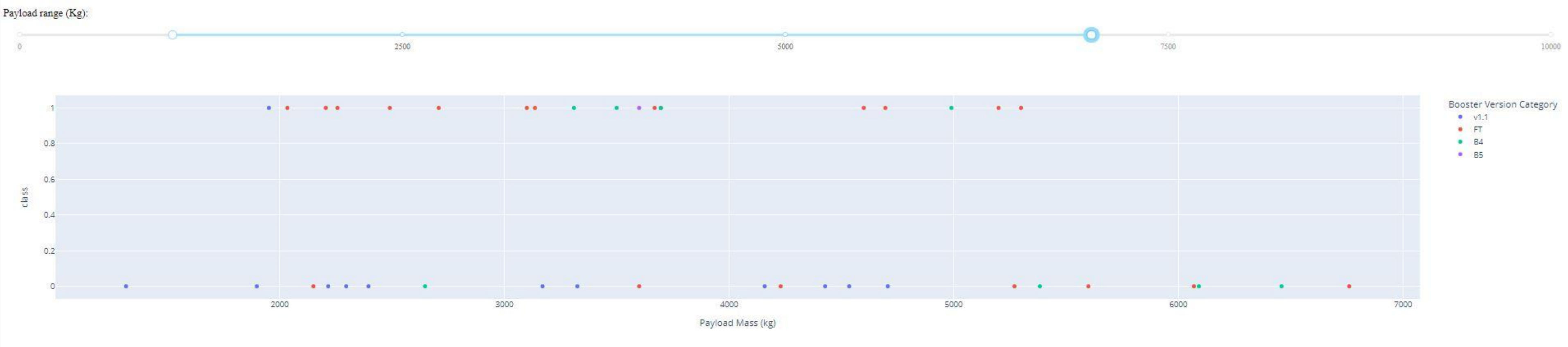
Success rate of CCAFS SLC-40 best ratio



CCAFS SLC-40

Here we can see that the success ratio of this site is 42.9%

Payload vs. Launch Outcome scatter plot for all sites



Here we can note that the payload range with the best success rate is between 1.500 and 3.800 kg

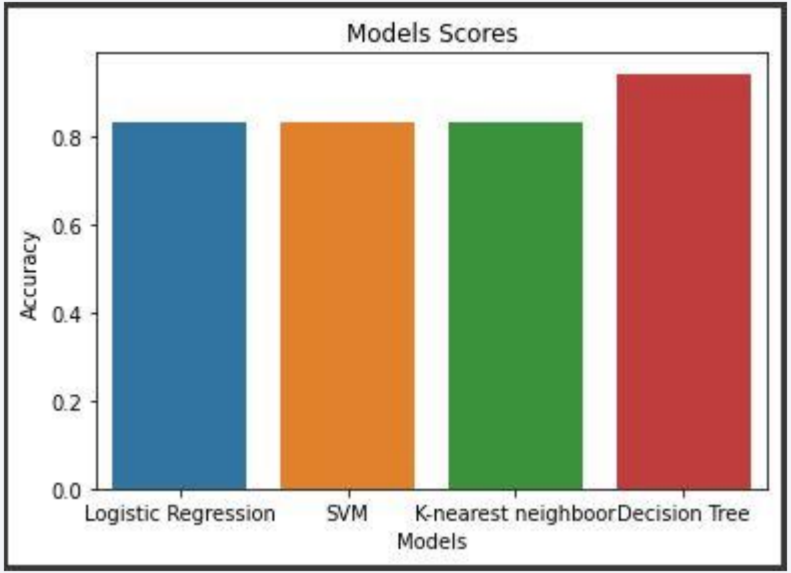
It is also clear that the FT Booster version is the best version

Section 6

Predictive Analysis (Classification)

Classification Accuracy

The best model based on the accuracy is a Decision Tree Classifier with a score of 0.9444

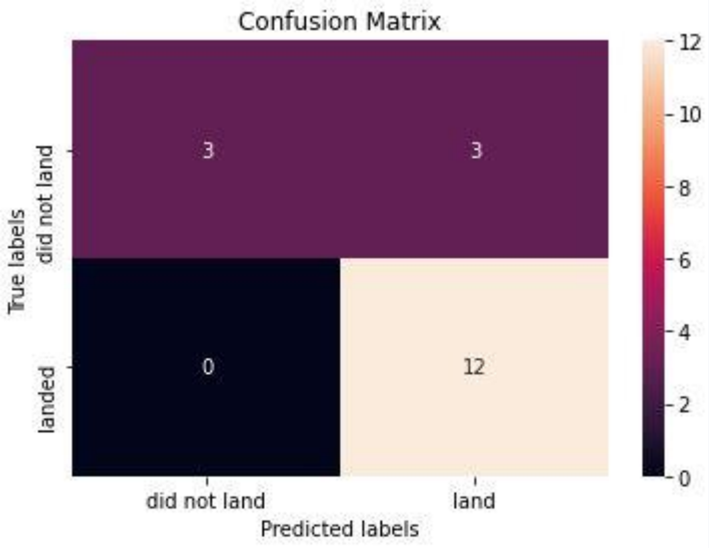


Confusion Matrix

The Confusion Matrix of the Decision Tree Classifier

True Positive : 12
False Negative : 0
True Negative : 3
False Positive : 3

The model is quite interesting as it predicts a lot of times the good labels, however 3 times it predicted the success of the mission and the mission failed. Reducing the amount of False Positive would be a good idea to avoid spending Millions and years of work. It could be done using Boosting or maybe look at a model with a lower accuracy but a better precision.



Conclusions

There are many parameters when considering launching rockets in space

- The Booster version is definitely one of this essential parameter
- The Orbit, Payload Mass are also important
- Machine Learning models can really helps to understand if a mission will be a success or a failure as it will learn from data of all previous launches. As we saw a model is able to predict with a high accuracy the reliability of a mission.
- However more data would be useful to have better, I have no doubt that engineer and scientists use these data in their predictions

Appendix

Thank you!